# 国際放送ニュース記事の自動文対応付け

西脇 正通　　熊野 正　　田中 英輝

ＮＨＫ放送技術研究所 人間情報科学

{ nishiwaki.m-hk, kumano.t-eq, tanaka.h-ja }@nhk.or.jp

## １．はじめに

筆者らは人手により多言語に翻訳された放送ニュース記事を、翻訳用例提示や機械翻訳に利用するために、自動文対応付けの研究を行っている。NHKの国際放送では日本語から直接翻訳される英語記事の他に、英語から 18 言語[i]の記事に翻訳している。

人手により翻訳した放送ニュース記事には、記事単位の対応関係が付与されている。しかし、意訳や要約、情報の追加・削除などが同時に行われており、記事対に含まれる文を先頭から順に並べるだけでは、文単位の対応が得られない。また、国際放送のために多様な言語に翻訳した放送ニュース記事を扱うためには、できるだけ言語に依存しない手法が必要となる。

本稿では、日本語から英語を経由して多言語に翻訳されたこれらの記事の特徴を調べ、人手により文対応の正解データを作成し、自動文対応付けの実験を行ったので報告する。なお、対象とした言語は 18 言語のうち、スペイン語、フランス語、アラビア語である。

## ２．放送ニュース記事の自動文対応付け

NHKの放送ニュース記事のうち、英語からスペイン語、フランス語、アラビア語の 3 言語すべてに翻訳されている記事の集合を、**4言語記事データ**として実験に使用した（表 1）。4言語記事データからは、単語の共起頻度を求めた。さらに、そのうち 20 組の記事を手法の評価に使用した。詳しくは後に説明する。

表 1　4言語記事データ

| | 英 | スペイン | フランス | アラビア |
|---|---|---|---|---|
| 記事数 | 4312 | | | |
| 平均文数 | 7.0 | 5.5 | 5.8 | 4.9 |
| 平均単語数 | 147 | 152 | 143 | 111 |
| 平均文字数 | 756 | 791 | 774 | 562 |

実験では、英語−スペイン語、英語−フランス語、英語−アラビア語、合計 3 種類の言語対で、対訳関係の文[ii]からなるセグメント（**対応組**：bead）[2][5]と、その並び（以下、**対応組列**と呼ぶ）を記事対ごとに推定し、結果を評価した。対応組を用いて文対応付けを行った例を図 1 に示す[iii]。

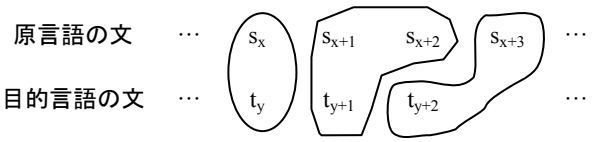


図 1 対応組を用いた文対応付けの例

## ３．正解データ

### ３．１．正解データの作成

自動文対応付けの手法を評価するために、4言語記事データからランダムに取り出した 20 組の記事について、人手で対応組列を作成し、文対応付けの正解データとした。対応組の種類（以下、**対応組種別**と呼ぶ）には、**削除**（1−0）[iv]、**挿入**（0−1）[v]や多対多も含み、対応数の上限は設けていない。

テキストを正確に比較して対応付けを行うためには、その言語対に精通した作業者が必要になる。しかし、作業者が異なると判断基準にゆれが生じる。そこで、各言語から日本語に人手でテキストを翻訳して、同一の作業者がこれを参照しながら文対応付けを行った。また、日本語テキストへは書かれている内容をなるべく過不足なく直訳調で翻訳し[vi]、文の対応付けをより正確に行えるようにした。

表 2 に作成した正解データの規模、表 3 に対応組列の具体例を示す。

表 2 正解データの規模

| 言語 | 英 | スペイン | フランス | アラビア |
|---|---|---|---|---|
| 記事数 | 20 | | | |
| 文数 | 128 | 105 | 115 | 103 |

[i] タイ、ベトナム、ビルマ、インドネシア、マレー、ベンガル、ヒンディ、ウルドゥ、ポルトガル、スペイン、イタリア、スウェーデン、ドイツ、ロシア、フランス、スワヒリ、ペルシャ、アラビア

[ii] 各記事のテキストは、文単位に分割されている。また、各言語の単語はスペースで区切られる。

[iii] 文はいずれかひとつの対応組に含まれる。異なる対応組に属した文どうしの前後関係は変わらない。

[iv] 翻訳されなかった文なので、目的言語側に対応する文がない状態

[v] 翻訳時に追加された文なので、原言語側に対応する文がない状態

[vi] 内容の過不足がない直訳調で翻訳するために、文単位でランダムに並べ替えたテキストを用いた。

表 3 英語-スペイン語の対応組列(正解データ)

| 対応組種別 | 英語 | スペイン語 |
|---|---|---|
| 1-1 | At least ten explosions occurred Sunday night in Baghdad, near an area that houses the headquarters of the US-led coalition.<br>(バグダッドで日曜日の夜、米主導の連合軍の本部がある地区の近くで少なくとも 10 回の爆発が起きた。) | Al menos diez explosiones se produjeron el domingo por la noche en Bagdad cerca de una zona que alberga la sede de la coalición liderada por Estados Unidos.<br>(日曜日の夜、バグダッドのアメリカを中心とした連合軍の本部を抱える地区の付近で、少なくとも 10 回の爆発が起きた。) |
| 1-1 | Sirens blared for several minutes.<br>(サイレンが数分間鳴り響いた。) | Como consecuencia, sonaron las sirenas durante unos minutos.<br>(その結果、数分間サイレンが鳴った。) |
| 2-1 | The US military says the explosions occurred near a convention center next to the headquarters.<br>(その爆発は本部の隣のコンベンション・センターの近くで起きた、と米軍は述べている。)<br><br>Iraq's new interim constitution is to be signed at the center as early as Monday.<br>(イラクの新しい暫定憲法は、早ければ月曜日にもセンターで署名されるはずである。) | El Ejército estadounidense dice que las explosiones tuvieron lugar cerca de un centro de convenciones próximo a la sede, donde hoy lunes se prevé la firma de la nueva Constitución interina de Iraq.<br>(アメリカ軍は、爆発は本部の近くにある会議場の付近で起きたと述べているが、そこは本日月曜日にイラクの新たな暫定憲法の署名が予定されていた場所である。) |

()内は正解データ作成のために翻訳した日本語テキスト

表 4 対応組種別の分布

| 対応組種別 | 英-スペイン | 英-フランス | 英-アラビア |
|---|---|---|---|
| 0-1 | 3.3% | 1.7% | 1.6% |
| 1-0 | 15% | 12% | 15% |
| 1-1 | 74% | 74% | 77% |
| 1-2 | 0.81% | 5.8% | 0.0% |
| 2-1 | 6.5% | 7.4% | 6.6% |
| 2-2 以上 | 0.0% | 0.0% | 0.0% |
| 合計 | 100% | 100% | 100% |
| 対応組数 | 123 | 121 | 122 |

### ３．２．国際放送ニュース記事の特徴

正解データから得た対応組種別の分布を表 4 に示す。この分布から、次のような特徴がわかった[vii]。

**① 1 対 1 対応(1-1)が最も多く、次に削除(1-0)が多い**
**② 交差・合成(2-2 以上)が見られない**

また、削除(1-0)された文の位置を観察すると、次の特徴があった[viii]。

**③ 1 文目と 2 文目は削除(1-0)されていない**

以上の特徴は、国際放送ニュース記事の特徴として、提案手法で利用する。

## ４．自動文対応付けの手法

### ４．１．Gale&Church(1991) の手法

対訳テキストの言語に依存しない文対応付けの手法のひとつとして、文字数の比を利用した手法 (以下、**Gale 手法**と呼ぶ) [3]がある。この手法は、次のように文対応を推定する。

**① 対応組種別を 6 種類、(0-1), (1-0), (1-1), (1-2), (2-1), (2-2) に限定する**
**② あらかじめ調べた対応組種別の生起確率と両言語の文字数の比から、対応組のコストを計算する[ix]**
**③ ダイナミック・プログラミングの手法で、最もコストの合計が小さい対応組列を解とする**

### ４．２．対応組の制限

それに対して提案手法では、3．2．節の国際放送ニュース記事の特徴を参考にして、以下のように文対応を推定した。

**① 対応組の種類は、3 種類(0-1, 1-0, 1-1)に限定する**
**② それぞれの記事対では、文の削除(1-0)と追加(0-1)を、同時には行わない**
**③ 1 文目は必ず対応付ける**

これらの条件を同時に満たす対応組列は、記事対で文の数を比べたとき、多い方の文数が m 少ない方の文数が n とすると、$_{(m-1)}C_{(n-1)}$通りの組合せとなる。例えば 9

[vii] 日本語から英語に一度翻訳されている記事の翻訳では、翻訳時に大きな編集を行うことが少なくなると考えられる。
[viii] 特に 1 文目はリード文として重要なので、削除されなかったと考えられる。
[ix] 文字数の比があらかじめ調べた平均に近く、対応組種別が出現数の多い(1-1)である場合、対応組のコストが低くなる。

文対 5 文のときでも 70 通りとなり、すべてを列挙しても十分計算ができる。そこで、対応組列を列挙し[x]、スコアが最も大きい対応組列を、自動文対応付けの解とした。

### ４．３．対応組列の評価

対応組列の評価には文字数の比を直接使わず[xi]、以下の式により対応組列のスコア(以下、**対応組列スコア**と呼ぶ)を求めた。

{対応組列スコア} :＝ {対応組スコア}の合計
{対応組スコア} :＝ {訳語対スコア}の合計

**訳語対スコア**は以下のようにして算出した。

① スペースで区切られた文字列を単語とする
② 訳語対は、単語の 1 対 1 対応とする(訳語関係の交差は可能)
③ 訳語対スコアには、単語共起の対数尤度比[4]を用いる
④ 訳語対スコアが大きな訳語対から、順に対応付けを確定する

## ５．実験

### ５．１．訳語対

訳語対には、4言語記事データでの出現回数が 2 回以上(表 5)の単語を使用した。ただし、出現回数が上位 100 以内に入る単語(表 6)は、その多くが冠詞や前置詞などの機能語なので除外した。そのため、表 7 に例で示した訳語対は、文対応付けに使用されなかった。また、対象テキスト全体で一回しか共起していない訳語対も、文対応付けに使用しなかった。

表 5 2 回以上出現した単語の異なり数

| 英 | スペイン | フランス | アラビア |
|---|---|---|---|
| 15,545 | 17,981 | 18,980 | 21,731 |

表 6 出現回数が上位 100 以内の単語(英語の例)

| 単語 | the | in | Japanese | say |
|---|---|---|---|---|
| 出現回数 | 4305 | 4114 | 1483 | 790 |

表 7 上位 100 以内の単語を含む訳語対の例

| スコア | 共起回数 | 英 | スペイン |
|---|---|---|---|
| 54139 | 4304 | the (4305) | de (4311) |
| 5029 | 3039 | of (4164) | un (3045) |
| 4850 | 1260 | Japan (1473) | Japón (1638) |
| 4582 | 2582 | for (2993) | del (3692) |

()内は出現回数

さらに、スコアの低い訳語対を除くことで、単語の対応付けの信頼性が上がると考えて、訳語対スコアが 100 以上の訳語対(表 8, 表 9)だけを使用した実験も行った。

表 8 訳語対の数

| | スペイン | フランス | アラビア |
|---|---|---|---|
| すべて | 2,764,833 | 2,759,096 | 2,621,570 |
| スコア 100 以上 | 4,837 | 4,411 | 3,945 |

表 9 スコアが高い訳語対の例

| スコア | 共起回数 | 英 | スペイン |
|---|---|---|---|
| 1451 | 329 | leader (419) | líder (421) |
| 1320 | 226 | Bush (300) | Bush (242) |
| 1304 | 233 | Bush (300) | George (264) |
| 775 | 161 | soldiers (192) | soldados (312) |
| 651 | 107 | system (132) | sistema (170) |

()内は出現回数

### ５．２．評価方法

評価には次の 2 種類の基準[xii]を採用し、正解率、再現率、F値をそれぞれ求めた。

**評価基準 A)**
**すべての種類の対応組を評価の対象として、対応組に含まれる両言語の文が、正解データと過不足なく一致する場合を、正解としてカウントする**

**評価基準 B)**
**(1-1)の対応組だけを評価の対象として、正解データと一致する(1-1)の対応組を、正解としてカウントする**

### ５．３．手法の比較

提案手法の評価では、ベースライン手法として、Gale 手法と、先頭から単純整列する手法(以下、**単純整列手法**と呼ぶ)の、2 つの手法と比較した。

単純整列手法は、記事の先頭から順に(1-1)の対応組として文を対応付け、片方の文が余ればそれを(0-1)または(1-0)の対応組とする。

正解データと同じ記事を対象に、文対応付けを行った結果、含まれる対応組の割合は、表 10、表 11のようになった。提案手法と単純整列手法は、含まれる対応組の割合が必ず同じになる[xiii]。

次に、各手法の評価結果は表 12 ～表 14 のようになった。

---

[x] 対応組の良さを計る指標では確実に正解を導き出せないので、部分的な最適解の積み重ねで解を得るダイナミック・プログラミングの手法よりも、可能な対応組列をすべて列挙して比較する手法の方が、適していると考えた。

[xi] 文の長さに関しては正規化を行っていないので、スコアの合計を求めるときに、間接的に文字数の比が関係する。

[xii] 文対応付けの結果を翻訳用例提示や機械翻訳に利用する場合、1 対 1 対応だけを利用することが多いので、評価基準 Bも採用した。そのほかに、(1-1), (1-2), (2-1), (2-2) の対応組だけを評価の対象とし、(1-2), (2-1), (2-2) の対応組は複数の 1 対 1 対応に分解してから、一致する 1 対 1 対応をカウントする方法でも評価したが、評価基準 B の評価と比べて、順序が入れ代わるような大きな違いはなかった。

[xiii] 提案手法と単純整列手法は、対応組の種類を(0-1), (1-0), (1-1) の 3 種類に限定しており、それぞれの記事対で、文の削除(1-0)と追加(0-1)を同時には行わないので、対応組種別の分布が同じになる。

表 10 対応組種別の分布(Gale 手法)

| 対応組 | 英-スペイン | 英-フランス | 英-アラビア |
|---|---|---|---|
| 0-1 | 0.0% | 0.0% | 0.0% |
| 1-0 | 0.0% | 0.0% | 0.0% |
| 1-1 | 72% | 77% | 72% |
| 1-2 | 1.0% | 5.5% | 0.99% |
| 2-1 | 24% | 17 % | 26% |
| 2-2 以上 | 3.0% | 0.0% | 0.99% |
| 合計 | 100% | 100% | 100% |
| 対応組数 | 100 | 109 | 101 |

表 11 対応組種別の分布(提案手法/単純整列手法)

| 対応組 | 英-スペイン | 英-フランス | 英-アラビア |
|---|---|---|---|
| 0-1 | 0.00% | 3.8% | 0.00% |
| 1-0 | 19% | 14% | 20% |
| 1-1 | 81% | 83% | 80% |
| 合計 | 100% | 100% | 100% |
| 対応組数 | 127 | 133 | 128 |

表 12 評価結果の比較(英-スペイン)

| 評価基準 | 評価値 | Gale 手法 | 単純整列手法 | 提案手法 |
|---|---|---|---|---|
| A | 正解率 | 62.0% | 61.4% | 76.6% |
| | 再現率 | 50.4% | 63.4% | 79.7% |
| | F 値 | 0.556 | 0.624 | 0.781 |
| B | 正解率 | 79.2% | 63.8% | 80.8% |
| | 再現率 | 62.6% | 73.6% | 92.3% |
| | F 値 | 0.699 | 0.684 | 0.862 |

表 13 評価結果の比較(英-フランス)

| 評価基準 | 評価値 | Gale 手法 | 単純整列手法 | 提案手法 |
|---|---|---|---|---|
| A | 正解率 | 70.6% | 48.1% | 70.6% |
| | 再現率 | 63.6% | 52.9% | 77.7% |
| | F 値 | 0.670 | 0.504 | 0.740 |
| B | 正解率 | 81.0% | 51.8% | 76.4% |
| | 再現率 | 76.4% | 64.0% | 94.4% |
| | F 値 | 0.786 | 0.573 | 0.844 |

表 14 評価結果の比較(英-アラビア)

| 評価基準 | 評価値 | Gale 手法 | 単純整列手法 | 提案手法 |
|---|---|---|---|---|
| A | 正解率 | 69.3% | 65.6% | 83.6% |
| | 再現率 | 57.4% | 68.9% | 87.7% |
| | F 値 | 0.628 | 0.672 | 0.856 |
| B | 正解率 | 89.0% | 70.9% | 87.4% |
| | 再現率 | 69.9% | 78.5% | 96.8% |
| | F 値 | 0.783 | 0.745 | 0.918 |

Gale 手法では、隣り合う(1-0)の対応と(1-1)の対応が、ひとつの(2-1)の対応として推定される傾向があった。

提案手法は、すべての言語対で、どちらの評価方法でも、F値が最も良かった。正解率、再現率も、ほとんどの場合で最も良かった。また、スコアが 100 以上の訳語対だけを、提案手法で使用した場合でも、結果は変わらなかった[xiv]。

出現回数が上位 100(表 6)以内の単語も訳語対に使用した場合は、各言語対でのF値が提案手法よりも、およそ 0.1 低下した。

## 6. まとめ

本稿では、放送ニュース記事が英語から 3 言語に翻訳されるときの特徴を分析し、その結果を元に自動文対応付け手法を提案した。提案手法と Gale&Church の手法、単純整列手法とを比較した結果、各言語対いずれにおいても提案手法の性能が他の 2 手法に比べて良いことがわかった。特に、1 対 1 対応を重視する評価基準 B では、英語-アラビア語が、F値 0.9 以上となった。

英語記事は、記者が作成した日本語記事からの翻訳結果である。この 1 回目の翻訳によって、表現方法や内容が調整されたため 2 回目の翻訳では 3.2. 節で示した特徴が現れたと考えられる。提案手法は、これらの特徴を利用した結果、対応組列の候補のスコアを総当たりで比較することが可能になった。また、特別なツールや辞書が不要であるため、新しい言語対に対して適用しやすい。

一方では、削除される文が必ずしも共通ではないことなどの違いも、特徴としてみられた。これは、翻訳した結果を同時に放送せず、放送する日時や長さが異なることも原因の一つだと考えられる。

今後は、それぞれの記事対で文の削除(1-0)と追加(0-1)を同時には行わないとする提案手法の制約をはずして、そのような記事対でも精度が低下しにくい手法に改善したい。また、オープンデータでの実験も行いたい。

参考文献


[1] Kay, M., R¨oscheisen, M.: Text-Translation Alignment. Computational Linguistics 19(1) (1993) 121–142

[2] Brown, P.F., Lai, J.C., Mercer, R.L.: Aligning Sentences in Parallel Corpora. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California (1991) 169–176

[3] Gale,W.A., Church, K.W.: A program for Aligning Sentences in Bilingual Corpora. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California (1991) 177–184

[4] Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19(1) (1993) 61-74.

[5] 長尾 真 編: 自然言語処理, 岩波講座ソフトウエア科学 15,岩波書店(1996).


[xiv] ただし、スコアが 1000 以上の訳語対だけを使用した場合は、訳語対の数が極端に減少するので、F値が低下した。